基于指数随机图模型的专利引用关系形成机制研究*
——以奈拉滨药物为例

杨冠灿[1] 刘占麟[2] 李纲[3]
[1] 中国人民大学信息资源管理学院 北京 100872 [2] 华盛顿大学工业工程系 西雅图 98105
[3] 武汉大学信息管理学院 武汉 430072

摘要:[目的/意义]专利引用关系形成问题是理解创新网络的一个重要问题。传统的回归模型对观察对象设定的独立性假设,无法将网络的结构效应因素整合到模型中来提供综合性的统计推断。指数随机图模型(ERGM,Exponential Random Graph Model)是一种创新性的统计推断方法,它能够将属性特征、自组织特征以及网络协同特征三种特征综合起来观察。[方法/过程]以奈拉滨药物的专利引文网络作为研究对象,利用ERGM系统检验了影响专利引用关系的五种机制:专利属性的主效应;专利引用时间的差值效应;专利引用关系的聚敛效应;专利引用关系的传递效应;专利引用关系的网络协同效应。[结果/结论]五种机制都在奈拉滨药物的专利引用关系的形成过程发挥了作用。但三种效应对于奈拉滨药物的专利引用关系的形成作用最为显著:共享发明人关系协同效应、共享家族关系协同效应、传递效应。一些辅助机制也会对专利引文关系形成产生影响,如引文时滞、权利要求数量和参考文献数量。

关键词:专利引用关系形成 指数随机图模型(ERGM) 奈拉滨 统计网络模型
分类号:G306
DOI:10.13266/j.issn.0252-3116.2019.10.009

1 引言

专利引文由于能够追踪技术发展的脉络,测量国家、区域间的技术扩散、技术溢出,衡量发明、技术的质量与价值,分析创新主体的技术战略行为[1-2],而在科技评价过程中具有十分重要的作用。近年来,学者们通过网络分析方法引入专利引文分析,涌现大量专利引文网络相关的研究成果,极大丰富了专利引文分析的视角,突破了传统单纯依赖专利引文频数进行分析的思路,采用可视化及描述性统计方法对专利引文的结构、动态特征开展了大量的讨论[3-4]。

然而,专利引用关系形成机制问题研究是目前研究中较为薄弱的一环,究其原因,主要表现为两点:①观察视角上的不足。专利引文网络的形成是一个复杂问题,其影响因素可能包括了专利引文网络自身演化过程,专利自身属性特征以及网络外部因素等;单纯地采用属性特征指标或者网络指标都难以很好的解释专利引用关系的形成机制问题[5];另外,很多在单一视角下成立的研究结论之间,在更高层次进行观察时可能存在冲突。②统计推断方法的不足。传统的统计推断方法,如回归方法,是以属性型数据为基础的,以独立性假设为前提的,而网络分析的核心对象是关系数据,因此,对其设定独立性假设是不合适的[6];同时,有一些专门针对网络数据的统计推断方法,如二次指派程序(Quadratic Assignment Procedure,QAP)方法虽然能符合网络数据的统计推断特点,但其受到了其框架的约束,在包容不同数据类型扩展性方面存在不足[7-8]。正是存在上面两点不足,需要进一步探索新的方法来回答专利引用关系形成的机制问题。

指数随机图模型(Exponential Random Graph Mod-

* 本文系国家自然科学基金项目"基于指数随机图模型的专利引用关系形成影响因素及机理研究"(项目编号:71403256)、国家自然科学基金项目"面向专利文本中实体关系抽取的远程监督方法研究"(项目编号:71704169)和国家自然科学基金重大项目"国家安全大数据综合信息集成与分析方法"(项目编号:71790612)研究成果之一。
作者简介:杨冠灿(ORCID:0000-0002-1706-1884),讲师,博士,E-mail:yanggc@ruc.edu.cn;刘占麟,博士研究生;李纲(ORCID:0000-0001-5573-6400),教授,博士,博士生导师。
收稿日期:2018-08-05 修回日期:2018-11-27 本文起止页码:75-86 本文责任编辑:杜杏叶

el,ERGM)是一种以关系形成(tie formation)为对象的研究方法[9],ERGM 是以关系数据为基础,以依赖性假设为条件,选择网络局部结构作为网络统计项来观察复杂网络的整体结构特征,从而获得对于网络复杂性、关联性以及随机性的整体认知的方法[8],因此,该方法能够克服在专利引用关系形成机制问题上研究所面临的上述两种不足,可以使研究人员获得对专利引用关系形成机制的更为全面的理解。本文的研究目标是:在关系形成理论的指引下,以可能对奈拉滨(Nelarabine)药物专利引文网络产生影响的主要机制为基础建立多个指数随机图模型,通过对各种机制对应的网络统计效应检验,帮助人们理解究竟哪些机制对于奈拉滨药物专利引文网络的形成产生了影响,影响效果如何。

文章依据如下顺序进行组织:第二部分是一个指数随机图建模的基本过程,简要描述影响专利引用关系形成存在的五种机制,以及如何转化为对应的局部网络配置(configuration)和网络统计项;第三部分介绍实验数据——奈拉滨药物的专利引文网络,以及利用统计方法直观展现上述五种机制相关统计特征的发现过程;第四部分是模型分析,包括模型比较、诊断以及拟合优度评价过程;第五部分则是结论与讨论,进一步讨论影响奈拉滨药物专利引文网络形成的五种核心机制对于专利引用关系形成的影响,回答哪些机制对于专利引用关系形成会产生影响,哪些机制的影响最为重要,以及对于未来药物研发的应用价值。

2 专利引用关系形成与 ERGM

2.1 影响专利引用关系形成的机制

学者们对于专利引用关系形成的机制问题已经开展了大量的研究,尤其是对于专利引文网络结构特征,产生了大量的研究成果[4,10-11],其中最有代表性的研究是 2017 年 A. B. Jaffe 教授对专利引文研究进展的一个梳理,她认为当前专利引文研究主要是从三个视角出发:测量发明的属性特征,如影响与原创性;追踪个体、机构、区域之间的知识流动;以及描绘创新网络图谱。如果从关系形成视角来理解上述研究,则可以将影响专利引用关系形成的因素归纳为三类:专利自身的属性、专利引文网络的自组织过程、以及引文网络受外部因素影响的过程[12]。本文从关系形成理论这个视角出发,通过梳理相关文献,提炼出五类影响专利引用关系形成的机制,这里所说的机制主要由两部分组成:包括影响专利引用关系形成的因素,以及这些因素对于专利引用关系形成的效应。需要说明的是这五类机制并不是排他的,未来可以根据研究的需求进行调整。

机制一:专利属性的主效应(main effects)。主效应主要是用来测量节点属性对于关系形成的影响。本研究中主要关注于两种专利属性特征,分别是专利权利要求项的数量以及专利参考文献的数量,相关研究认为专利权利要求反映了技术排他权的边界,而参考文献则呈现了专利对现有技术的依赖程度[2]。目前,相关研究已经证明了权利要求项数量对于专利被引频次具有正向影响作用[13],同样,专利参考文献数量对于专利被引频次也具有正向影响作用[4]。与标准统计分析不同,ERGM 模型关注节点对之间的关系,因此,主效应所测量的统计量是专利对属性的汇总值,而非单个专利的属性值。

机制二:专利引用时间的差值效应(difference effects)。以往的研究证明了专利引文具有队列效应(cohort effect),即专利被引的数量随着时间增长而增长。专利引文时滞常被用于测量技术的技术生命周期,解释创新的速度或者技术发展的速度,相关研究显示,专利有引用较新专利的倾向,表现在引文时滞上也就是说专利引文时滞间隔短的专利更易于形成专利引用关系[14-15]。具体到网络效应上,专利引文时滞可以表现为:专利引用对所对应的授权年之间差值对专利引用关系形成的影响。

机制三:专利引用关系的聚敛效应(activity effect)。前向引文数量(被引频次)由于在一定程度上反映了该专利后续的技术影响力,一直以来都是研究关注的焦点[16-17]。对应到引文网络中,专利引用关系聚敛效应是对前向引文数量分布网络结构层面的刻画,它将高被引的专利视为具有星状结构的网络局部配置(从中心节点链接入两条或者多条弧),从而观察这种配置对于网络关系形成的影响,如“富人俱乐部”[18]或者“倾向链接”[19]现象,上述问题正是聚敛效应要测量的内容。因此,机制三是指专利对之间形成聚敛结构对专利引用关系形成的影响。

机制四:专利引用关系的传递效应(transitivity effect)。传递效应主要观察的是一种特殊的网络局部配置——传递闭合(transitivity closure)。传递闭合是一种纯网络结构特征,早期研究主要是通过聚集系数等指标对其进行测量的。在引文网络中,传递闭合表现为两个方面特点:一方面,传递闭合是在 2-路径构造基础增加的一条弧,该弧的增加使得“遗失链接”显

性化，构造内部的关系更为稳健，该特征可以用于分析专利技术的演化路径[20]；另一方面，传递闭合构造中，度分布并不均匀，某些节点具有更多的入度，而这种在传递闭合构造中的入度优势要优于单纯聚敛效应构造中的入度优势，于是，传递效应也可用于识别知识流动过程中的源头[21-22]。因此，机制四是指专利对之间形成传递结构对专利引用关系形成的影响。

机制五：专利引文网络的网络协同效应(Covariates effect)。与上述机制不同，网络协同效应不是指专利引文网络内部的网络结构特征，而是以其他网络与专利引文网络之间的协同特征为观察对象的。现有相关研究揭示了专利权人、专利发明人地理位置上的临近[23]与专利引用关系形成之间有相关关系，H. D. White 的研究进一步确认专利引文网络实际是由两种网络结构特征共同作用的结果，即社会交流结构(social structure)和技术交流结构(intellectual structure)[24]，当然，专利间文本的语义相似性也能在一定程度上影响专利引用关系的形成[25]。因此，机制五是指专利对之间其他网络关系对专利引用关系形成的影响。

2.2 影响机制到网络局部构造

ERGM 是一种以关系形成为对象的研究方法，其发轫于 1959 年 P. Erdos 和 A. Renyi 提出的社会网络统计分析模型，1996 年 S. Wasserman 将上述模型扩展成为可以包含图中任何统计配置的 ERGM/p* 模型，1999 年 J. Anderson 提出了对上述模型的参数化估计方法使得模型有了重要的进展[26]。ERGM 是一个可以根据研究内容进行调整的扩展模型，其最一般的形式为：

$$\Pr(Y=y)=\left(\frac{1}{\kappa}\right)exp\{\sum_{A}\eta_A g_A(y)\} \quad 公式(1)$$

其中，求和是包含所有的配置 A 的加总，η_A 是对应的配置 A 的参数，该参数可以用来判定观测网络中特定网络统计量的影响力，$g_A(y)=\prod_{y_{ij}\in A}y_{ij}$是对应配置的网络统计量，$\kappa$ 是标准化常数，确保公式为适当的概率分布[27]。简单说来，ERGM 模型的核心任务就是给具有某些特定机制组合的网络赋予权值的过程。因此，上式也可以写成一种条件 Logit 的形式：

$$\mathrm{Logit}(P(Y_{ij}=1 \mid n\ patent, Y_{ij}^{C}))=\sum_{A}\eta_A\,\delta g_A(y) \quad 公式(2)$$

其中，Y_{ij}^{C} 表示网络中除 Y_{ij}之外的其他链接关系，而 $\delta g_A(y)$则表示当链接 Y_{ij}从 0 到 1 变化时 g_A 的变化值，因此公式(2)的含义是在网络中其他连线已经确定条件下，预测一条新的连线出现的概率。

ERGM 是在对有序的局部网络配置进行观察基础上的建模，通过特定的参数估计过程，局部网络配置所对应的参数值可以被计算出来，从而实现对于复杂网络结构的统计推断过程。ERGM 从理论上解决了传统方法无法对复杂网络条件下混合变量(同时包含多个属性变量与关系变量)的评价问题，能够在全网层次上解释专利引用关系的成因，因此，就有可能做出更准确的预测。表 1 展示了如何将影响专利引用关系产生了五种机制转化为可计量的网络统计项的过程。

3 数据来源与探索

3.1 数据来源

本研究关注的是一种小分子创新抗癌药物，其药物的中文名是奈拉滨(Nelarabine)，其在美国上市的商品名为 Arranon(阿仑恩)。之所以选择奈拉滨是出于如下考虑：

药物研发过程会经历多个过程，一个完整的生命周期通常 15-20 年左右，因此，一些药物自身的特点，如早期药物研发阶段的技术转移、临床 I、II、III 期的审查结果、药物潜在的适用症范围、药物商业化阶段的转移(融资、并购)、商品化后期毒性、药效方面的负面报道以及药物的更新换代、专利到期导致的"专利悬崖"等都可能潜在影响某个药物相关专利的规模与引文网络的特征[28-30]，也会使得 ERGM 模型发现的引用关系形成的特点可能会存在偏差(bias)。因此，在药物案例选择时我们希望尽可能从简单药物入手逐步深入，这里，我们考虑了筛选一个简单药物的三个条件：从研发到生产环节相对简单，尽可能少涉及合并、转移；药物的适应症范围较窄，负面毒性、药效报道相对较少；为了要更全面观察专利引文网络，观察期要截止到药物核心专利到期后一段时期。

该药物的早期药物研发阶段主要是美国国家癌症研究所和 Glaxo Wellcome 合作开展的，该项研发方 Glaxo Wellcome 和 SmithKline Beecham 与 2000 年合并形成 GlaxoSmithKline(即葛兰素史克公司)。虽然在早期药物研发阶段，该药物研发活动中存在多主体参与的现象，但由于核心专利均是在 2005 年之后产生的，引文也是在 2006 年之后出现的，因此，早期药物研发活动中的多参与主体特征并不会对本研究中的引文关系形成产生较大影响[31]。另外，虽 2015 年以后奈拉滨药物研发及生产团队由诺华收购，但由于是整体收购，该药物的研发、生产环节仍完全由原葛兰素史克公司团队所控制，因此，也不会对最终引文关系形成产生较大影响。

表 1　影响专利引用关系形成五种机制(结构效应)对应的网络构造表

机制(影响因素)	机制(待检验效应)	参数	网络配置图示	统计项计算公式
专利对(施引)权利要求项数量和越大越可能会产生引用关系	主效应	Nodeicov (claims)		$\sum_{i,j} x_{ij}(y_i + y_j)$
专利引用时滞为 n 年的专利之间越可能产生引用	差值效应	Absdiffcat (year. n)		$\sum_{i,j} x_{ij}\lvert y_i - y_j\rvert$
专利对之间入度分布对专利引用关系形成的影响	聚敛效应	Gwidegree	k	$e^{\alpha}\sum_{i=1}^{n-1}\{1-(1-e^{-\alpha})^{i}\}(\sum_{i} x_{+i})$
专利对之间的传递闭合结构对于专利引用关系形成的影响	传递效应	Gwesp	k	$e^{\alpha}\sum_{i=1}^{n-2}\{1-(1-e^{-\alpha})^{i}\}(\sum_{i<j} x_{ij}\sum_{k\neq i,j} x_{ik}x_{kj})$
共享专利家族关系的专利之间越可能会产生引用	网络协同效应	Edgecov (famnet)		$\sum_{i,j} x_{ij}y_{ij}$

奈拉滨药物于 2005 年 10 月被 FDA 批准上市，是经过 FDA 特殊审批流程的孤儿药(即用于治疗罕见疾病的药物)[32]，该药物的适应症是：用于治疗至少两种治疗方案无效或治疗后复发的 T 细胞急性淋巴细胞性白血病(T-ALL)和 T 细胞淋巴母细胞性淋巴瘤(T-LBL)[33]，该药物潜在的适用症范围有限。根据相关文献，它是治疗 T 细胞恶性肿瘤的有效药物，各期临床试验均取得了较好的效果，主要面临的问题是需要通过调整剂量来控制神经毒性的风险。后期的研究主要是集中在组合用药上，根据当前的研究尚未出现对于该药物完全替代性新药[34]。

利用 PubChem 化合物结构数据库(PubChem Compound Database)进行检索，检索策略选择奈拉滨药物的 PubChem CID 是 3011155 (https://pubchem.ncbi.nlm.nih.gov/compound/3011155)，可以获得关于该药物两方面的专利信息，首先，是核心专利信息，主要是 FDA 橙皮书中公开的核心专利信息，另外，该数据库也提供一个根据化合物结构式在专利全文中识别出的相关专利信息[35]，截至 2017 年 12 月 31 日，检索结果显示为 3 035 条专利相关文献。在数据预处理环节，本研究限定为 1998 年 – 2016 年美国专利授权数据之间的引用关系，最终，数据集中包含涉及奈拉滨药物化合物的 1 165 项美国专利授权以及 1 168 条专利引用关系。数据补充环节，主要采用 PatentsView 专利数据库(http://www.patentsview.org/api/doc.html)以及美国专利局(USPTO)授权专利数据库全文与图像数据库(http://patft.upsto.gov)进行数据补充。经数据补充后，数据集被进一步加工为网络数据格式，其由两个数据集构成，专利属性数据与专利间关系数据。

(1)专利属性数据。专利属性数据包含了 4 个字段，其中，Patent_id 是专利数据的标识符，其他三个字段是分别是与该专利相关的三个属性信息，分别是专利授权年、专利权利要求项数量以及专利参考文献的数量。出于数据标准化的考虑对权利要求项以及参考文献数量分别进行了处理。具体处理方式参考表 2。

表 2　属性数据统计项及其解释

统计项	名称	解释	最小值	最大值	平均值
patent_id	授权专利号码	授权专利的号码	5424295	9527925	–
year	专利授权年	专利授权年	1995	2016	2013
claims	专利权利要求项数量(sqrt2)	专利权利要求项数量的平方根	1	9	3.70
references	专利参考文献数量(sqrt4)	专利参考文献数量的四次方根	0	5	2.23

(2)专利间关系数据。专利关系数据(见表 3)包含了 5 个字段，其中，patent_id_ego 和 patent_id_alter 分别表示关系的两端，这里由于专利引用关系是有向关系，因此，将其他类型的关系均转化为有向关系进行处理。关系数据中包含三种关系，分别是专利之间共享申请人关系、专利之间的引用关系以及专利之间共享专利家族的关系。

3.2　数据分析

在统计建模之前，利用图形可视化和描述性统计方法对数据进行观察是非常有必要的。相关研究发现：真实网络往往与随机网络之间存在许多结构性差异，这些差异能够帮助我们将真实网络与简单随机网络区分开来。经过基本的数据探索，我们发现奈拉滨

杨冠灿，刘占麟，李纲. 基于指数随机图模型的专利引用关系形成机制研究——以奈拉滨药物为例[J]. 图书情报工作，2019,63(10):75－86.

表 3　关系数据统计项及其解释

统计项	名称	解释	关系数量
patent_id_ego	授权专利号码(链出)	链出的专利号码	－
patent_id_alter	授权专利号码(链入)	链入的专利号码	－
rel_inventor	共享发明人关系	如果链入与链出的专利号码之间至少包含一项共同的专利发明人,就认为他们具备共享发明人关系	10 643
rel_citing	专利引用关系	授权专利之间的引用关系	1 168
rel_family	共享专利家族关系	如果链入与链出的专利号码之间至少包含一项共同的专利家族信息,就认为他们是共享专利家族关系	472

药物专利引文网络在上述五种机制上均表现出于随机网络不同的网络结构效应：

(1)专利属性的主效应特征。表 4 展示了专利引用对之间各自对应的专利权利要求项数量,并以此建立一个混淆矩阵(Confusion matrix),即针对具有不同专利权利要求数量的专利引用对各种可能组合的形式进行统计,检验专利引用对在引用关系形成上是否受到了专利权利要求项数量属性特征的影响。在表 4 中,列表示代表的是专利引用对中的施引方,而行则是代表了专利引用对中的被引方。不难观察到在该混淆矩阵中,左上部分矩阵块中(行 1－5 与列 1－5)的数据密度更高,该特征似乎说明:权利要求项数量较少的专利之间建立引用关系的概率高。同时,我们进一步观察,发现在表 4 的左上部分矩阵块中,对角线的上三角形区域较对角线下三角形区域的数据密度明显更高,该特征可能说明,权利要求项数量偏低的专利更有可能被引用。当然,这个结论还需要通过模型进行检验。同时,对专利参考文献进行观察时也发现存在类似的主效应特征。

表 4　专利引用对之间基于专利权利要求项数量的混淆矩阵

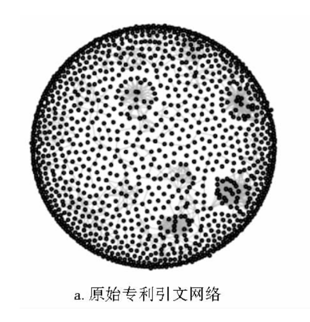
施引＼被引		专利权利要求项数量									
		1	2	3	4	5	6	7	8	9	行汇总
专利权利要求项数量	1	23	41	49	14	12	6	0	2	0	147
	2	27	49	79	40	17	17	6	4	1	240
	3	16	34	61	53	24	22	10	6	2	228
	4	11	17	52	118	52	22	17	6	0	295
	5	9	23	24	32	36	13	9	3	1	150
	6	5	9	8	15	5	15	6	9	1	73
	7	1	0	5	9	7	3	3	2	0	30
	8	0	0	1	1	1	0	1	0	0	4
	9	0	0	1	0	0	0	0	0	0	1
	列汇总	92	173	280	282	154	98	52	32	5	1 168

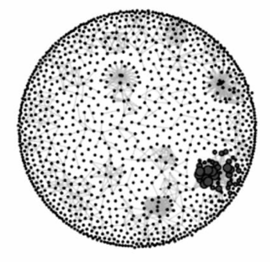
(2)专利引用对的时序特征。由于整个引文跨度较长,但奈拉滨药物引文网络早期施引与被引都较少,因此,为了更显著的表现引文时滞的特点,表 5 中仅截选了 2007 年至 2016 年间的专利引用关系进行展示(引用关系数量为 1 076)。通过建立施引专利与被引专利之间基于授权时间的混淆矩阵,能够比较清晰的观察专利授权专利之间在时间上的特征。通过对表 5 的观察,我们可以发现两点:该药物的专利施引(表 5 行汇总)是从 2013 年之后开始爆发,2012 年的专利施引量仅为 26,而 2013 年后专利施引量 113,说明,2013 年以后该药物逐步成为药物研发领域的关注热点;其次,围绕在邻接矩阵对角线区域的 2－5 年范围存在一个高密集区域,该密集区域可能说明专利引文的时滞存在一个间隔期的偏好,即专利引用关系更倾向在授权时间间隔为 2－5 年范围内的专利之间发生。

表 5　专利引用对之间基于专利授权年份的混淆矩阵(2007 年－2016 年)

施引＼被引		专利的授权年份										
		2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	行汇总
专利的授权年份	2007	0	－	－	－	－	－	－	－	－	－	0
	2008	1	0	－	－	－	－	－	－	－	－	1
	2009	6	2	2	－	－	－	－	－	－	－	10
	2010	1	2	2	1	－	－	－	－	－	－	6
	2011	3	5	10	8	0	－	－	－	－	－	26
	2012	1	2	7	5	10	2	－	－	－	－	27
	2013	2	13	9	21	36	26	6	－	－	－	113
	2014	2	6	9	13	19	25	37	1	－	－	112
	2015	1	13	8	21	33	42	125	47	10	－	300
	2016	4	7	7	13	36	57	104	121	92	40	481
	列汇总	21	50	54	82	134	152	272	169	102	40	1 076

(3)专利引文网络呈现整体稀疏与局部聚集特征。首先,网络的密度仅为0.000 867 说明该网络是一个整体较为稀疏的网络。图 1a 展现了原始的奈拉滨药物专利引文网络整体稀疏的特征,即并不存在高度聚敛的中心节点,部分高密度区域的影响范围有限。图 1b 和图 1c 展现的是经过权威得分(Authority)与枢纽得分(Hub)算法[36]计算后,对网络中的节点大小进行缩放后的图像[37]。通过比较图 1 的三幅图像,我们观察到在网络的局部的高密度区域中,部分专利间的施引与被引非常频繁,存在局部聚集特征。


a. 原始专利引文网络


b. 枢纽得分加权专利引文网络

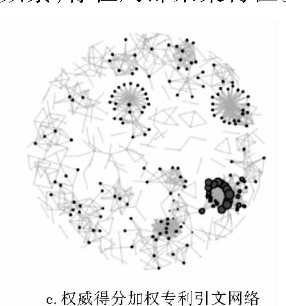
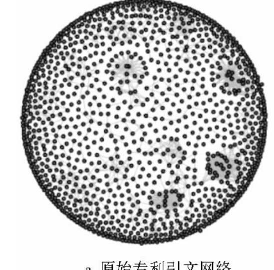
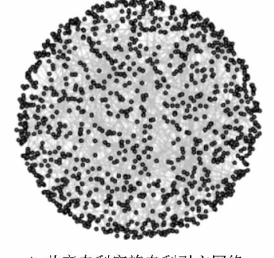

c. 权威得分加权专利引文网络

图 1 奈拉滨药物专利引文网络局部聚集特征展示

(4)网络协同机制。首先,如图 2 所示,我们比较了三种专利引文网络:专利引用网络、共享发明人的专利引文网络、以及共享家族关系的专利引文网络。从图 2 中,不难观察到三种专利引文网络的一些基本特征,尤其是图 2c 中所示的共享发明人的专利引文网络展示了一个紧密连接的核心成分(Component),说明在奈拉滨药物专利引文网络中存在一个存在高度自引倾向的"小圈子",在这个"小圈子"内的任何存在引用关系的专利引用对至少有一个发明人是相同的(即两篇专利共享发明人的关系)。由于高度自引特征一定程度上反映了该药物技术发展对现有技术的依赖程度,以及核心研发团队对于专利引用关系形成具有重要影响,于是,这种高度自引特征就成为后续统计推断过程需要重要关注的内容。

相对于共享发明人专利引文网络呈现出较为清晰的核心成分,共享专利家族关系的专利引文网络则显得非常的杂乱,从图形 2b 几乎无法发现任何网络结构特征。但当进一步计算三种网络的自相关矩阵时(见表 6),专利引文网络与共享专利家族关系的专利引文网络之间存在 0.301 的自相关关系,与此同时,专利引文网络与共享发明人专利引文网络之间仅存在 0.193 的自相关关系;尤其是三个网络中关系数量的分布并不是均匀的,共享发明人专利引文网络的关系数量是 10 643条;共享专利家族专利引文网络的关系仅为 472 条。当将上述两条信息结合起来考虑,不难想象,专利引文网络与共享专利家族关系的专利引文网络之间是存在某种高度协同性的,而这种协同性与结构特征无关,可能暗示存在某种强规则或业务逻辑对专利引文形成产生了影响,当然,上述判断也需要依据网络统计推断进行确认。


a. 原始专利引文网络


b. 共享专利家族专利引文网络

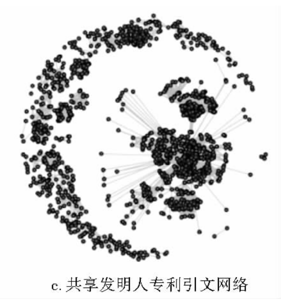

c. 共享发明人专利引文网络

图 2 三种专利引文网络协同效应展示

表 6 三种网络关系的自相关矩阵

	共享发明人专利引文网络	专利引用关系	共享家族关系专利引文网络
共享发明人专利引文网络	1	0.193	0.161
专利引文网络	0.193	1	0.301
共享家族关系专利引文网络	0.161	0.301	1

另一个值得一提的问题:如何在数据探索阶段发现网络协同效应的存在?发现哪些网络与专利引用关系网络具有协同作用并不是一个非常直观的过程。一个经验是我们在早期针对专利引文属性特征进行探索时,发现一些现有文献中非常显著的属性特征,如专利

发明人数量、专利家族规模的效果并不理想，因此，我们并不是直接采用专利引用对之间在属性特征是否存在主效应、差值、同质或异质作为统计特征，而是通过观察引用关系对之间是否共享发明人或者贡献专利家族关系来构建网络，然后，测量上述两个网络与专利引用关系网络之间的协同性。通过这个转化过程很多时候可以获得意想不到的效果。

4 结果分析

4.1 参数估计

模型评价是贯穿整个建模过程的一个重要环节，通常而言，ERG 模型的研究过程是：首先将零模型（随机生成的网络）作为基线模型，然后，逐步增加不同机制对应的网络统计项形成新的模型，并利用 ERG 模型对上述模型进行参数估计，最终，对多个模型的结果进行诊断、拟合优度评价、比较与解释。本文选用 R 的 statnet 包对表 7 中的各项模型进行参数估计，其中，零模型、主效应模型、差值模型以及协同关系模型均是采用的是最大似然估计方法进行参数估计，而对于几何加权模型则是采用马尔可夫链蒙特卡罗极大似然估计法（MCMC MLE）[38]。

表 7 是五种模型的统计摘要表。通过对五种模型统计摘要的比较，尤其是对网络的参数估计值及其统计显著性的分析，可以获得对网络统计项的初步统计观察。“专利对（施引）权利要求项数量和”在全部模型中均显示为显著且负向，说明当其他条件不变的情况下，在专利对（施引）权利要求项数量和越大，他们之间建立引用关系的概率就越小；“专利对（施引）参考文献数量和”在全部模型中均显示为显著且正向，说明当其他条件不变的情况下，在专利对（施引）参考文献数量和越大，他们之间建立引用关系的概率就越大；同时，需要注意的是：“专利对（被引）参考文献数量和”在除几何加权模型外的其他模型中均显示为显著且正向，可能的解释是当模型加入几何加权入度分布或几何加权边共享伙伴统计项后，可能在上述三种因素之间存在某种程度的相关关系。差值模型中的前两项“引文时滞（2 年）”“引文时滞（3 年）”在全部模型中均显示为显著且正向，说明专利对之间如果授权时间之间间隔不超过 3 年，那么他们之间建立引用关系的概率就越大；值得关注的是后两项“引文时滞（4 年）”“引文时滞（5 年）”，当模型加入几何加权入度分布或几何加权边共享伙伴统计项后，“引文时滞（4 年）”则不显著了，可能的解释是几何加权入度分布或几何加权边共享伙伴统计项与引文引文时滞（4 年）之间存在相关因素。“共享专利家族关系”与“共享发明人关系”在协同关系模型和几何加权模型下都呈现为显著且正向，说明当其他条件不变的情况下，在专利对之间如果存在“共享专利家族关系”或者“共享发明人关系”，那么，他们之间建立引用关系的概率就越大，另外值得注意的是，“共享专利家族关系”与“共享发明人关系”的参数值非常高，分别为（2.8，2.9）以及（6.34，5.41）这说明这两项网络协同机制对于建立引用关系具有非常大的正向影响。最后，“几何加权入度分布”为显著负向，专利节点对之间建立引用关系的概率要小于随机发生引用关系的概率，但“几何加权边共享伙伴”则为显著正向，专利对之间建立引用关系的概率要大于随机发生引用关系的概率，看起来似乎矛盾，但综合起来实际上进一步说明网络结构上整体稀疏与局部聚集特征并存的现象。

4.2 模型诊断

模型诊断（model diagnostics）能够辅助判断估计算法是否已经收敛还是存在近似退化问题，进而判断究竟是模型本身还是模型评价设置条件需要进行调整[39]。图 3 展示几何加权模型部分统计项在模型最后迭代阶段呈现的状态。在图 3 左侧的绘图，以模型中的每一个统计项为单位，利用 MCMC 链作一个时间序列来展示统计项的变化情况，图 3 右侧的绘图则显示了对应 MCMC 链的分布图。如果模型能够收敛，模型中每一个统计项的图表将会表现为以 0 为中心随机变化，这里 0 代表观测网络对应统计项的统计值。在几何加权项模型中，大多数统计项的图表都是围绕 0 随机变化的，因此，模型诊断的结果显示几何加权项模型是一个稳定的模型。

4.3 模型拟合

虽然，在参数估计环节一些网络统计项已经表现出了统计上的显著性，并且反映出了一些与前期根据探索性分析所观察出的模式一致的特征，对于模型的效度已经进行了初步的检验，但还需要更为系统地的检验：究竟仿真模型能够在多大程度上反映观察网络的结构特征。下面，我们将从两个方面对模型的拟合优度进行评价：

（1）利用 AIC 和 BIC 统计结果进行拟合优度的评价。AIC 和 BIC 方法是基于对数似然估计结果的，即观测网络中 Yij（真实发生的联系）概率与 Yij 的期望概率之间的差异。根据表 7，零模型的 AIC 是 18 806，主效应模型的 AIC 是 17 640，较之前的零模型有较大

表 7　零模型、主效应模型、差值模型、协同关系模型、几何加权模型的统计摘要表

效应类别	网络统计项	名称(statnet)	参数估计值(SE)				
			零模型	主效应模型	差值模型	协同关系模型	几何加权模型
	弧	Arc	-7.049 0.02 ***	-9.751 0.14 ***	-10.158 0.14 ***	-11.761 0.16 ***	-9.463 0.14 ***
主效应	专利对(施引)权利要求项数量和	Nodeocov(claims)	-	-0.248 0.01 ***	-0.245 0.01 ***	-0.171 0.02 ***	-0.092 0.01 *
	专利对(被引)权利要求项数量和	Nodeicov (claims)	-	-0.014 0.01	-	-	-
	专利对(施引)参考文献数量和	Nodeocov (references)	-	0.482 0.03 ***	0.473 0.03 ***	0.265 0.04 ***	0.064 0.03 ***
	专利对(被引)参考文献数量和	Nodeicov (references)	-	0.950 0.03 ***	0.947 0.03 ***	0.899 0.04 ***	-
差值效应	引用时滞(2 年)	Absdiffcat (year.2)	-	-	0.823 0.07 ***	0.852 0.08 ***	0.58 0.08 ***
	引用时滞(3 年)	Absdiffcat (year.3)	-	-	0.570 0.08 ***	0.746 0.09 ***	0.476 0.09 ***
	引用时滞(4 年)	Absdiffcat (year.4)	-	-	0.450 0.10 ***	0.619 0.11 ***	-
	引用时滞(5 年)	Absdiffcat (year.5)	-	-	0.594 0.11 ***	0.972 0.13 ***	-
协同效应	共享专利家族关系	Edgecov (famnet)	-	-	-	2.80 0.11 ***	2.936 0.10 ***
	共享发明人关系	Edgecov(invnet)	-	-	-	6.34 0.08 ***	5.411 0.09 ***
聚敛效应	几何加权入度分布	Gwidegree (α = 0.3)	-	-	-	-	-1.827 0.11 ***
传递效应	几何加权边共享伙伴	Gwesp (α = 0.3)	-	-	-	-	0.68 0.05 ***
拟合优度	赤池信息准则	AIC	18 806	17 640	17 510	8 515	7 970
	贝叶斯信息标准	BIC	18 818	17 701	17 607	8 636	8 080

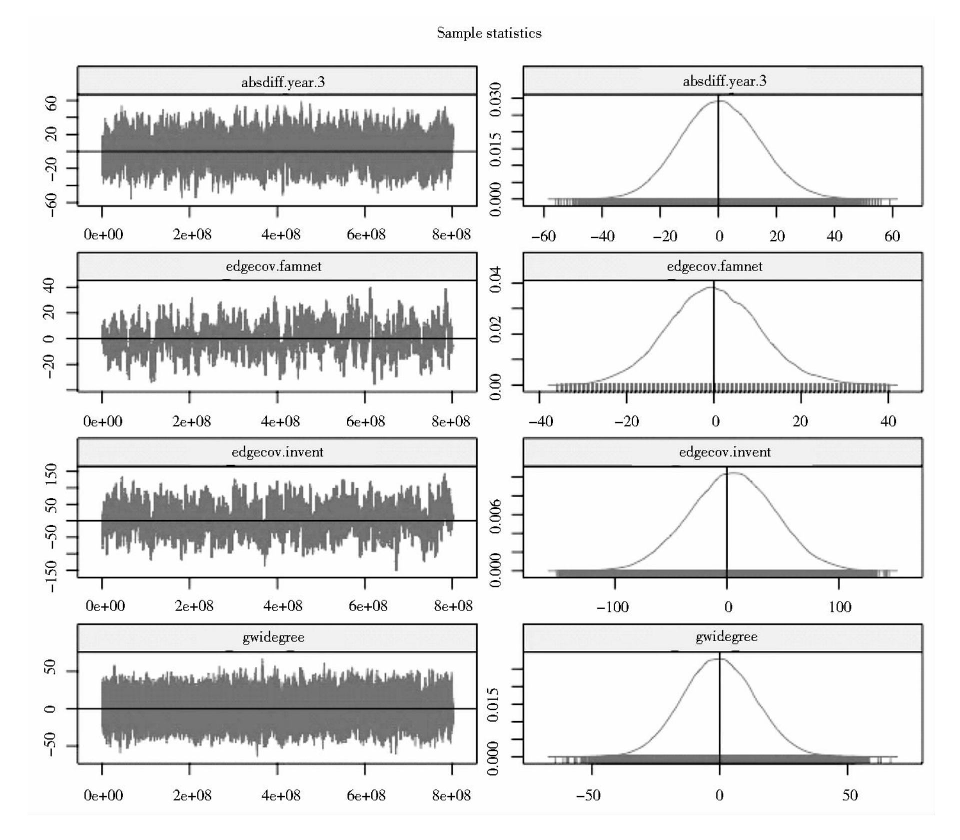


图 3　几何加权模型图形化模型诊断结果(部分)

提升；差值模型的 AIC 是 17 510 虽然也有所下降，但较之前的主效应模型改变并不明显；协同关系模型较之前的差值模型有了一个显著的降幅，AIC 下降到了 8 515，说明协同关系模型中的两种机制共享专利家族与共享发明人关系的协同作用对应 ERG 模型拟合优度的改进具有重要的作用。

（2）然而，AIC 与 BIC 等方法是适合于以独立性假设为基础的观测数据的，但当模型更加复杂，例如几何加权模型增加了依赖性统计项时，就需要采用基于仿真的模型拟合优度评价方法。拟合优度评价的过程也可以采用可视化图形观察的方法，当限定其他网络特征不变的前提下，比较观测网络中每一个参数的对数优势比以及仿真网络中对数优势比的范围。图 4 的组图是针对几何加权模型仿真网络进行拟合优度评价的结果。其中，黑色线代表专利引文网络的观测结果；灰色线以及箱型图则代表了仿真网络在 95% 的置信区间时的测量结果。当黑色线落在灰色线条之间时，说明仿真网络能够很好的代表真实的专利引文网络的结构特征。因此，图 4 的组图说明，仿真网络基本上能够拟合真实网络的四种结构特征（入度中心度、出度中心度、边共享伙伴以及二元组共享伙伴），但在边共享伙伴这一特征上和真实网络还有一定差异。

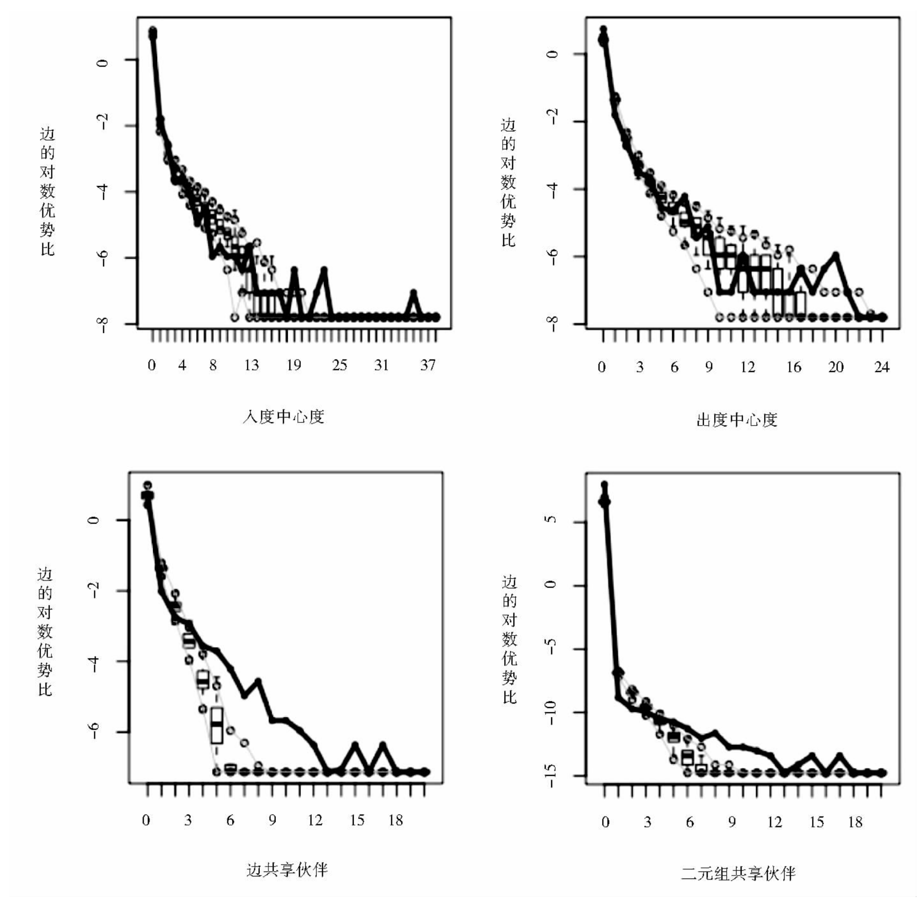


图 4　几何加权模型拟合优度评价的图形化观察比较

4.4　模型解释

本研究通过比较五种模型（零模型、主效应模型、差值模型、协同关系模型、几何加权模型）的多个统计结果以及图形化拟合优度指标，我们发现几何加权模型具有最佳的网络仿真效果。在模型构建的过程中，我们观察到模型对于网络仿真效果改进最大的地方有三处：①增加协同效应的统计项，即将专利之间共享专利家族关系的网络以及共发明人网络视为网络协变量作为统计项，用于预测专利之间建立引用关系的概率；②增加主效应特征统计项，即考虑属性因素对于专利之间建立引用关系概率的影响，包括专利权利要求项数量以及专利参考文献数量；③增加几何加权统计项，即考虑入度分布，边共享伙伴对于专利之间建立引用关系概率的影响。

具体而言，从 ERG 模型拟合优度改进的效果来解读，“共享发明人关系”统计项对于专利引用关系形成的影响是最大的。“共享发明人关系”统计项所代表的专利引用关系协同效应类似于文献间“作者自引”

效应，说明奈拉滨药物专利引文网络是围绕一个存在高度自引倾向的小圈子展开的，该小圈子在网络中同时占据了枢纽和权威的位置（参见图 1 和图 2），于是，可以说该小圈子的发展决定了整个奈拉滨药物专利引文网络的形态。

“共享专利家族关系”统计项展现了对专利引用关系形成起到重要的另一个重要维度，即专利申请背后的业务逻辑，即专利申请者会利用衍生专利申请来进行专利布局，如采取围栏策略，用以扩大专利的保护周期。这种规则是一种强业务规则，我们无法从网络结构特征中窥见端倪（参见图 2b 和表 6），但却深刻的影响到了专利引用关系的形成。

“几何加权边共享伙伴”所代表的专利引用关系的传递效应类似于“朋友的朋友也是朋友”。对于引文网络而言，传递效应的存在并不难理解，更值得关注的点在于在几何加权模型中，由于增加了“几何加权边共享伙伴”统计项，网络中其他统计项的相对影响作用出现了下降的趋势，说明“几何加权边共享伙伴”统计项对某些统计项存在一定程度的替代作用。这一点恰恰是指数随机图模型的优势所在能够分析存在复杂嵌套关系的多个变量给出统计推断，这一点是传统回归模型无法胜任的。

当然，专利属性的主效应机制也是存在一定作用的。例如，“专利对（施引）权利要求项数量和”与“专利对（施引）权利要求项数量和”两项专利属性的主效应机制说明，在奈拉滨药物专利引文网络中，专利更倾向于引用哪些权利要求项数量较少的专利，即采用主动避开竞争对手的权利要求范围的策略[30]；同时，专利更倾向于引用哪些参考文献数量较多的专利，即采用主动信息披露策略，避免因信息披露不全导致在后期诉讼环节处于不利地位[14]。

专利引用时间的差值效应也是存在一定作用的。“引用时滞（2 年）”和“引用时滞（3 年）”都表现出显著的差值效应，这一点在专利引文中非常常见，但需要注意的是“引用时滞（4 年以上）”统计项则在加入专利引用关系的聚敛效应、传递效应机制后不再显著了，这表明网络结构特征（如聚敛效应）与差值效应之间存在一定的替代作用。例如，如果三篇专利之间存在引用关系构成了一个传递三元组，那么三元组中两篇专利之间既存在直接引用关系也存在间接引用关系，这种情况下，引用时滞通常会比仅存在直接引用关系的专利对要长。合理的解释是虽然专利引文网络中存在一部分专利对之间的引用时滞较长的现象，但这些引用对之间往往也同时存在传递三元组结构，因此，引用时滞（4 年以上）统计项对在考虑了传递性之后就不再显著了。

5 结论与不足

本文尝试使用了一种新的统计推断方法——指数随机图模型方法，该方法为本研究提供了独特视角，使得本研究能够对复杂网络条件下混合变量进行综合评价，从而能在更广泛的层次上解释专利引用关系的形成问题。在微观结构特征设计方面，本文考虑了五种机制：主效应、差值效应、协同效应、聚敛效应、传递效应，这五种机制涵盖了网络内部自组织结构特征、外部网络协同作用以及专利内部属性特征，这些特征存在多重关系与高度嵌套的局部结构，是传统以独立性假设为前提的回归模型难以胜任的。

研究的主要结论如下：就奈拉滨药物引文网络而言，专利引用关系形成主要是受到了三方面的影响：由共享发明人关系与专利引用关系之间的协同效应显示存在一个具有高度自引倾向的“小圈子”，这个小圈子很大程度上影响了整个奈拉滨药物研发的方向；共享专利家族关系与专利引用关系之间的协同效应显示专利申请背后的业务逻辑的作用——利用专利家族进行布局；专利引文网络内部自组织网络特征——如传递性，显示专利关系的形成并不是一个随机过程，而是对于既有网络结构有着较强的依赖性。

同时，在奈拉滨药物专利引文网络中，还有一些影响专利引用关系形成的辅助因素也非常值得关注：网络结构特征（如传递效应）对引文时滞（4 年以上）的替代作用；专利更倾向于引用权利要求项数量较少的专利，即采用主动避开竞争对手的权利要求范围的策略；同时，专利更倾向于引用参考文献数量较多的专利，即采用主动信息披露策略，避免因信息披露不全导致在后期诉讼环节处于不利地位。

本文存在一定的研究不足值得未来研究继续深入探讨。首先，本研究是针对一个特定的药物的，因此，我们目前所发现的所有对于专利引用关系形成的解释仅适用于奈拉滨药物专利引文网络，不具有普适性，这种局限性是由指数随机图模型生成模型的特点所决定的。但如果我们能够同时对多个药物进行分析，则仍有可能归纳出一些对于专利引用关系形成普适性的解释。针对多药物的比较研究未来会是我们研究的一个重要方向。其次，我们只是从一个截面数据条件来观察专利引用关系形成，未来的研究中，从一个动态视角

来观察专利引用关系形成的影响也是我们研究的一个主要方向。

参考文献：

[1] OECD. OECD patent statistics manual[M]. Paris:OECD Publishing, 2009.

[2] JAFFE A B, DE RASSENFOSSE G. Patent citation data in social science research: overview and best practices[J]. Journal of the Association for Information Science and Technology, 2017, 68(6): 1360－1374.

[3] YANG G C, LI G, LI C Y, Using the comprehensive patent citation network (CPC) to evaluate patent value[J]. Scientometrics, 2015, 105(3): 1319－1346.

[4] VAN RAAN A F J. Patent citations analysis and its value in research evaluation: a review and a new approach to map technology-relevant research[J]. Journal of data and information science, 2017, 2(1): 545－538.

[5] MORRIS S A, VAN DER VEER MARTENS B. Mapping research specialties[J]. Annual review of information science and technology, 2008, 42(1): 213－295.

[6] ARRIETA PAREDES M P, CRONIN B. Exponential random graph models for management research: a case study of executive recruitment[J]. European management journal, 2017, 35(3): 373－382.

[7] ROSE KIM J Y, HOWARD M, COX PAHNKE E. Understanding network formation in strategy research: exponential random graph models[J]. Strategic management journal, 2016, 37(1): 22－44.

[8] GOODREAU S M, HANDCOCK M S, BUTTS C T. Statnet: software tools for the representation, visualization, analysis and simulation of network data[J]. Journal of statistical software, 2008, 24(1): 1－11.

[9] ROBINS G, PATTISON P, KALISH Y. An introduction to exponential random graph (p*) models for social networks[J]. Social networks, 2007, 29(2): 173－191.

[10] JAFFE A B, TRAJTENBERG M. Patents, citations, and innovations[M]. New York: MIT Press, 2002.

[11] ALCÁCER J, GITTELMAN M. Patent citations as a measure of knowledge flows: the influence of examiner citations[J]. Review of economics and statistics, 2006, 88(4): 774－779.

[12] ROBINS G. Doing social network research[M]. London: SAGE, 2015.

[13] FISCHER T, LEIDINGER J. Testing patent value indicators on directly observed patent value-an empirical analysis of Ocean Tomo patent auctions[J]. Research policy, 2014, 43(3): 519－529.

[14] ALCÁCER J, GITTELMAN M, SAMPAT B. Applicant and examiner citations in U.S. patents: an overview and analysis[J]. Research policy, 2009, 38(2): 415－427.

[15] HALL B H, JAFFE A B, TRAJTENBERG M. The NBER patent citation data file: lessons, insights and methodological tools[R]. Cambridge: National Bureau of Economic Research, 2001.

[16] BENSON C L, MAGEE C L. Quantitative determination of technological improvement from patent data[J]. Public library of science, 2015, 10(4): e0121635.

[17] CZARNITZKI D, HUSSINGER K, SCHNEIDER C. "Wacky" patents meet economic indicators[J]. Economics letters, 2011, 113(2): 131－134.

[18] SMILKOV D, KOCAREV L. Rich-club and page-club coefficients for directed graphs[J]. Physica a: statistical mechanics and its applications, 2010, 389(11): 2290－2299.

[19] BRANTLE T F, FALLAH M H. Complex innovation networks, patent citations and power laws[C]//PICMET '07－2007 Portland international conference on management of engineering & technology. Portland: IEEE,2007: 540－549.

[20] WANG J C, CHIANG C H, LIN S W. Network structure of innovation: can brokerage or closure predict patent quality? [J]. Scientometrics, springer netherlands, 2010, 84(3): 735－748.

[21] BATAGELJ V. Efficient algorithms for citation network analysis [EB/OL]. [2017－12－31]. https://arxiv. org/abs/cs/0309023. pdf.

[22] HUNG S W, WANG A P. Examining the small world phenomenon in the patent citation network: a case study of the radio frequency identification (RFID) network[J]. Scientometrics, 2009, 82(1): 121－134.

[23] ALMEIDA P, KOGUT B. The exploration of technological diversity and geographic localization in innovation: start-up firms in the semiconductor industry[J]. Small business economics, 1997, 9(1): 21－31.

[24] WHITE H D, WELLMAN B, NAZER N. Does citation reflect social structure?: Longitudinal evidence from the "Globenet" interdisciplinary research group[J]. Journal of the Association for Information Science and Technology, 2004, 55(2): 111－126.

[25] YAN E, DING Y. Scholarly network similarities: how bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other[J]. Journal of the Association for Information Science and Technology, 2012, 63(7): 1313－1326.

[26] SNIJDERS T A B, PATTISON P E, ROBINS G L. New specifications for exponential random graph models[J]. Sociological methodology, 2006, 36(1): 99－153.

[27] ROBINS G, SNIJDERS T, WANG P. Recent developments in exponential random graph (p*) models for social networks[J]. Social networks, 2007, 29(2): 192－215.

[28] THORNE N, AULD D S, INGLESE J. Apparent activity in high-throughput screening: origins of compound-dependent assay interference[J]. Current opinion in chemical biology, 2010, 14(3): 315－324.

[29] LIM S Y, SUH M. Intellectual property business models using pa-

tent acquisition: a case study of royalty pharma inc[J]. Journal of commercial biotechnology, 2016, 22(2): 6 – 18.

[30] WAGNER S, WAKEMAN S. What do patent-based measures tell us about product commercialization? evidence from the pharmaceutical industry[J]. Research policy, 2016, 45(5): 1091 – 1102.

[31] KISOR D F. Collaboration to meet a therapeutic need: the development of nelarabine[J/OL]. Clinical nedicine. 2009, 1: 1317 – 1320. [2018 – 12 – 06]. https://dio.org/10.4137/CMT.s2909.

[32] FDA approval for nelarabine[EB/OL]. [2017 – 12 – 06]. https://www.cancer.gov/about-cancer/treatment/drugs/fda-nelarabine.

[33] COHEN M H, JOHNSON J R, JUSTICE R. FDA drug approval summary: nelarabine (Arranon) for the treatment of T-cell lymphoblastic leukemia/lymphoma[J]. The oncologist, 2008, 13(6): 709 – 714.

[34] KADIA T M, GANDHI V. Nelarabine in the treatment of pediatric and adult patients with T-cell acute lymphoblastic leukemia and lymphoma[J]. Expert review of hematology, 2016, 10(1): 1 – 8.

[35] PAPADATOS G, DAVIES M, DEDMAN N. SureChEMBL: a large-scale, chemically annotated patent document database[J]. Nucleic acids research, 2016, 44(D1): D1220 – D1228.

[36] MARRA M, EMROUZNEJAD A, HO W. The value of indirect ties in citation networks: SNA analysis with OWA operator weights[J]. Information sciences, 2015, 314: 135 – 151.

[37] LUKE D. A user's guide to network analysis in R[M]. Cham: Springer International Publishing, 2015.

[38] DUBNJAKOVIC A. An evaluation of exponential random graph modeling and its use in library and information science studies[J]. Library & information science research, 2016, 38(3): 259 – 264.

[39] ROBINS G, PATTISON P, WANG P. Closure, connectivity and degree distributions: exponential random graph (p*) models for directed social networks[J]. Social networks, 2009, 31(2): 105 – 117.

作者贡献说明:

杨冠灿:负责论文思路框架构建,主体内容撰写、实验与结果分析;

刘占麟:进行数据探索、代码调优、实验结果分析;

李纲:确定论文思路,提供论文修改建议。

Understanding Mechanisms of Patent Citation Formation Based on ERGM: A Case Study of the Nelarabine Drug

Yang Guancan[1] Liu Zhanlin[2] Li Gang[3]

[1] School of Information Resource Management of Renmin University of China, Beijing 1000872

[2] Department of Industrial and Systems Engineering, University of Washington, Seattle 98105

[3] School of Information Management of Wuhan University, Wuhan 430072

Abstract: [**Purpose/significance**] The Formation of patent citation is necessary to understand innovation networks. The independence assumption set by the Conventional regression model for observed objects cannot integrate the structural effect factors of the network into the model to provide comprehensive statistical inference. ERGMs (exponential random graph model) represent a methodological innovation of statistical inference for networks given their ability to model actor attributes along with endogenous self-organizational processes and exogenous network covariates. [**Method/process**] In this paper, ERGMs are applied to systematic inspect the five mechanisms affecting patent citation formation in a sample of Nelarabine drug. The five mechanisms contain main effect, difference effect of citation lag, and activity effect, transitivity effect and network covariates. [**Result/conclusion**] We find that five different types of mechanisms play diverse roles in patent citation formation. And three of effects among these mechanisms have significant impacts on citation formation of nelarabine drug: network covariates based on shared inventors and shared patent family membership, and transitivity effect. In addition, some aided mechanism play a supporting role on patent citation formation, such as difference of time lag, main effects of number of claims and reference.

Keywords: patent citations formation ERG (exponential random graph) nelarabine drug discovery statistical network analysis